Weight-Based Firefly Algorithm for Document Clustering

Athraa Jasim Mohammed[1], Yuhanis Yusof[2], and Husniza Husni[2],

School of Computing, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
[1]autoathraa@yahoo.com
[2]{yuhanis, husniza}@uum.edu.my

Abstract. Existing clustering techniques have many drawbacks and this includes being trapped in a local optima. In this paper, we introduce the utilization of a new meta-heuristics algorithm, namely the Firefly algorithm (FA) to increase solution diversity. FA is a nature-inspired algorithm that is used in many optimization problems. The FA is realized in document clustering by executing it on Reuters-21578 database. The algorithm identifies documents that has the highest light intensity in a search space and represents it as a centroid. This is followed by recognizing similar documents using the cosine similarity function. Documents that are similar to the centroid are located into one cluster and dissimilar in the other. Experiments performed on the chosen dataset produce high values of Purity and F-measure. Hence, suggesting that the proposed Firefly algorithm is a possible approach in document clustering.

Keywords: Firefly algorithm, partitional clustering, hierarchical clustering, text clustering.

1 Introduction

Clustering is a process of grouping documents into a cluster. Similar documents are grouped in the same cluster and dissimilar documents in another cluster [1]. Furthermore, it is an unsupervised learning that does not require pre-defined classes for the intended documents. In general clustering algorithms can be classified into two categories; hierarchical clustering and partition clustering algorithms [2, 3].

Hierarchical clustering algorithm is a technique to build a hierarchy of clusters. There are two approaches of this technique [1]. The first approach is agglomerative hierarchical clustering which started by working from bottom to top, meaning that every object in a single cluster is merged based on similarity between clusters [4]. The second approach is the divisive hierarchical clustering which operates from top to bottom. In this approach, objects in cluster are separated using one of the partition clustering techniques. In undertaking a hierarchical clustering, one does not require to determine the number of output clusters [5]. On the other hand, Partition clustering returns an unstructured set of clusters. Partition techniques have some drawbacks; they require a pre-defined number of cluster and an initial cluster centers. This paper

T. Herawan et al. (eds.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7_30,

uses Weight-based Firefly Algorithm (WFA) in a divisive hierarchical clustering to overcome such problems. The proposed WFA attempts to overcome the problem of trapping in local optima (of identifying the k number of clusters) by using Firefly Algorithm (FA). The total weight of document is assigned as the initial light intensity of a firefly. Furthermore, the attractiveness between documents is later undertaken based on Euclidean distance. The proposed approach then finds the center of cluster that produces the highest brightness.

The remainder of the paper is structured as below: In section 2, we provide related work. Section 3, present the proposed Weight-based Firefly Algorithm (WFA) approach. The evaluation is presented in section 4, followed by the conclusion in section 5.

2 Related Works

Text clustering is a useful technique for organizing text documents as clusters, the similar text is in one group and dissimilar text is in another group [6]. One of the most famous clustering techniques is the K-means which is classified as a type of partitioning clustering. Another well-known algorithm is the Principal Direction Divisive Partitioning (PDDP) for divisive hierarchical clustering [7]. Divisive clustering is one type of hierarchical clustering that it is used to construct a hierarchy of clusters. One drawback of divisive clustering is its low performance, hence, leading to the combination with Intelligent Swarm methods. Particle Swarm Optimization algorithm is one type of Intelligent Swarm methods that it is integrated with divisive clustering approach [8]. The experiment result indicates high performance and robustness with lower running time.

The K-means algorithm has been widely utilized in the domain of clustering. Nevertheless, due to its random initial centroids, work has been reported to fall into local optima. Such situation has led researchers to integrate K-means with optimization techniques such as Particle Swarm Optimization algorithm (PSO) [9]. However, the result of the proposed sequential approach was no better than having PSO as an individual clustering method. Despite such result, it was learned that for the Wine and Iris [10] datasets, the combination of K-means and PSO generate better results.

Another type of Intelligent Swarm methods is the Firefly Algorithm (FA). Firefly algorithm was developed by Xin-She Yang in 2007 at Cambridge University. It has two important issues, the light intensity and the attractiveness. For optimization problems, the light intensity, I, of a Firefly at a particular location, x, can be determined by $I(x) \alpha f(x)$. The attractiveness β is relative. It changes depends on the distance between two fireflies [11]. Firefly algorithm is utilized in many optimization problems such as image processing which it is used to search for multiple thresholds [12]. Furthermore, the performance of Firefly algorithm was also studied in numerical clustering [13]. The objective function of such work was to minimize the distance between a center and the documents. The result was efficient, robust, and reliable that generates optimal cluster centers compared with two nature inspired algorithms;

Particle Swarm Optimization (PSO) and Artificial Bee Colony (ABC). The proposed FA solved the local optima problem but the number of k cluster was pre-defined.

There has also been effort in integrating FA and K-means for data clustering. The FA was employed to find the center of clusters while the K-means was utilized in finding its clusters [14]. The proposed hybrid KFA minimized the intra-cluster distance and reduced the clustering error compared to the K-means, PSO and KPSO. However, this hybrid approach was implemented only on numerical data sets. In addition, a hybrid of Firefly Algorithm and K-harmonic Means algorithm was also undertaken on numerical data sets [15]. In this paper, we propose a clustering technique utilizes FA to identify the initial cluster center using an objective function that is formulated based on the term frequency of a document. The FA will also identify documents that will be grouped into the identified centroid.

3 Method

Theproposed Weight-based Firefly Algorithm consists of the following steps:

3.1 Data Preprocessing

The data preprocessing is an important phase in web mining as it extract various information from websites and represent it as a database. This phase includes seven steps and they are follows. First, extraction of important text is done, the tags that contains title and body. Second, the selected text is cleaned from digit and special characters. Each text from documents that was cleaned is split to words. All of the words in each document are analyzed for its length. If the words length is less than two then it will be removed because it is useless in search process otherwise remain. Fifth, the stop words is removed from the list of words and this includes words like the, on, in, etc. The sixth step is on utilizing a stemmer algorithm that transforms a word into its root. Lastly, the word frequency is calculated based on its occurrence in the document [17].

3.2 Development of Vector Space Model

Vector space model VSM has been widely used to represent data in document clustering. Each document is represented as a vector in the vector space [6]. The VSM includes several steps: In the first step, a term-frequency (TF) database is created. The rows include terms (words) and the columns represent the documents. The intersection between row and column contain the occurrence of each terms (term frequency) [16]. Secondly, a normalized matrix is created that where the occurrence of terms is normalized between (0, 1) through calculate the length of each document by using equation 1 [17]:

$$Length = \sqrt{\sum_{i=1}^{m} V_i(d)^2} \quad (1)$$

Where, m is the number of term in a collection, V is the term frequency, d is the document. Then, normalized term frequency is divided on document length by using equation 2:

$$EN = \frac{TF}{Length} \quad (2)$$

Where, TF is term frequency. Then the weighting matrix tf-idf is created to find the weight of each term in the document [17]. In order to do so, we need to calculate the inverse documents frequency idf by using equation 3:

$$idf = \log N/dft \quad (3)$$

Where, N is the total number of documents in the collection, dft is the number of documents in the collection that contain a term. Following to that, we determine the weight of a term by using equation 4 [17]:

$$tf - idf_{t,d} = tf_{t,d} * idf_t \quad (4)$$

The total weighted of each document is obtained using equation 5:

$$total\ weight_{d_j} = \sum_{i=1}^{m} tf - idf_{t_i,d_j} \quad (5)$$

Where, j is the number of documents, i is the number of the terms.

3.3 Text Clustering

Firefly algorithm has two important features, (a): the light intensity and (b): the attractiveness. In an optimization problems, the light intensity I of a firefly at a particular location x can be determined using objective function f(x). The attractiveness β is relative. It changes depending on the distance between two fireflies. The attractiveness β formula is shown in equation 6 [11]:

$$\beta = \beta_0 exp^{(-\gamma r_{ij}^2)} \quad (6)$$

The movement of one document i to another document j is determined based on equation 7:

$$X^{i} = X^{i} + \beta * (X^{j} - X^{i}) + \alpha \quad (7)$$

Where, Xi is the position of first document; Xj is the position of second document in training data set.

In this paper, we propose that each document is represented by a single firefly and the total weight of the document is the initial brightness I of the firefly. The highest brightness is indicated by the highest total weight value and represents the best point. The best point hence indicates the center of a cluster. The distance between two documents is then calculated using Euclidean distance function [14] as shown in equation 8:

$$Euclidean\ distance(X_i, X_j) = \sqrt[2]{(X_i - X_j)^2} \quad (8)$$

The WFA is used to determine document in the collection that has the highest brightness and is later used as centroid of cluster. Once this is done, we then find the similar documents for the centroid using cosine similarity matrix. Documents having high similarity value is located in the first cluster while the ones with lower values in a second cluster. Such an approach requires threshold. The cosine intra-similarity is defined in equation 9 [5]:

$$Intra - sim(C_i) = \sum_{j=1}^{m} (X_j * V_j) \quad (9)$$

Where, Ci is the output cluster, j is the number of terms in the collection, Xj is the documents in cluster C, Vj is the center of cluster.

The second cluster that contains documents with low similarity value against the centroid will again enter the text clustering phase (weight-based firefly algorithm to find new centroids). The process of finding a centroid and its cluster continues until it reaches the last document. The proposed Weight-based Firefly Algorithm is shown in Figure 1.

```
Generate Initial population of firefly randomly x^i where i=1, 2, .., n,
n=number of fireflies (documents).
Initial Light Intensity, I=total weight of document.
Define light absorption coefficient γ, initial γ=1
Define the randomization parameter α, α=0.7
Define initial attractiveness β0 = 1.0
While t < N
For i=1 to N
For j=1 to N
If (total weight Ii < total weight Ij) {
Calculate distance between i, j using equation 8.
Calculate attractiveness using equation 6.
Move document i to j using equation 7.
Update light intensity I^i = I^i + β
End For j
End Fori
Loop
Rank to find best document.
```

Fig.1. The proposed Weight-based Firefly Algorithm (WFA)

4 Evaluation

The proposed WFA is tested on a standard text classification dataset which is the Reuters-21578 [18]. In the undertaken experiment, data is divided into three sets; two parts are used for training and one part is used for testing. The training data is used to build the clusters of documents and the test data is used to measure the clustering quality of the proposed algorithm. Two data collection from Reuters-21578 was chosen which are the RE0 and RE1. Description of the chosen documents is presented in Table I [19].

Table I. Description of Data

Data Set	No. of Documents	Classes	No. of Training data	No. of Testing data	No. of Terms
RE0	201	13	134	67	2149
RE1	192	25	128	64	2156

As for the evaluation, the Classification Error Percentage (CEP) [12], Purity and F-measure [19] are used as performance measurement. CEP is calculated by counting the number of documents that is wrongly classified. It is shown in equation 10 [13]:

$$CEP = \frac{number\ of\ wrong\ classified\ documents}{total\ number\ of\ documents\ in\ test\ data\ set} * 100 \quad (10)$$

Purity on the other hand is a measure of clustering quality [19]. The purity depends on the maximum number of documents in class Ωk and in cluster Cj respectively. The equation is in 11 [19]:

$$P(\Omega_k, C_j) = \mathrm{Max_k} \, |\Omega_k \cap C_j| \quad (11)$$

The cluster purity calculated as in equation 12[19]:

$$Purity = \sum_{\Omega_k \in \{\Omega_1, \ldots, \Omega_C\}} \frac{P(\Omega_k, C_j)}{N} \quad (12)$$

To measure the accuracy, the F-measure [19] is employed and it depends on the recall and precision values [20]. The total F-measure is the summation average of F-measure for all class. The equation to collect maximum value of F-measure is in equation 13 [19]:

$$F(\Omega_k) = \max_{C_j \in \{C_1, \ldots, C_k\}} \left(\frac{2 * R(\Omega_k, C_j) * P(\Omega_k, C_j)}{R(\Omega_k, C_j) + P(\Omega_k, C_j)} \right) \quad (13)$$

Where: $R(\Omega_k, C_j)$is recall measure and $P(\Omega_k, C_j)$is precision measure. The equation for total F-measure is in equation 14 [19]:

$$Total\ F-measure = -\sum_{k=1}^{c} \frac{|\Omega_k|}{N} * \max(F(\Omega_k)) \tag{14}$$

The CEP, Purity and F-measure results are shown in Table II. Form the table, it is noted that the first dataset, RE0, has CEP of 23.9, purity of 0.7089 and F-measure of 0.5535. As for the RE1, the CEP is equal to 21.9, while its purity and F-measure are 0.7890 and 0.5768 respectively. Based on literature, it is learned that a good clustering is when the CEP value is low and the F-measure and Purity values are high [19]. Hence, the obtained result of CEP which is less than 30, and Purity and F-measure values higher than 0.5 indicates that the Firefly algorithm produces good clusters.

Table II. Result of WFA

Data Sets	CEP	Purity	F-measure
RE0	23.880	0.7089	0.5535
RE1	21.875	0.7890	0.5768

5 Conclusion

A new approach for document clustering is presented using a meta-heuristics algorithm which is the Firefly. In this paper, we propose that each document is represented by a single firefly and the total weight of a document is the initial brightness of the firefly. The point (document) with the highest brightness is later identified as the centroid. Such an operation is assumed to be a new approach in utilizing Firefly in document clustering. Such an approach operates by defining that the significance of total weight of documents is equal to the light intensity in firefly. In theory, the firefly which has the highest light will attract other fireflies. Hence, in this work, the proposed WFA uses the highest total weight of document to represent the centroid and attract other documents based on similarity between centroids and documents.The performance of the proposed WFA is tested on a standard text classification dataset which is the Reuters-21578 and is evaluated using three performance measurements which are the Classification Error Percentage (CEP), Purity and F-measure. The obtained results indicated that the proposed Weight-based Firefly Algorithm would become a competitor in the area of data clustering.

Additionally, the proposed WFA can be operationalized in the form of a search engine. It could be used to optimize the organization of index file structures into clusters. Hence, may lead to a better precision and recall of a search engine and reduces its computational time.

References

1. Das, S., Abraham, A., Konar, A.: Metaheuristic Clustering, Springer, Heidelberg (2009).
2. AnithaElavarasi, S., Akilandeswari, J., Sathiyabma, B.: A survay on Partition Clustering Algorithms. In: International journal of Enterprise Computing and Business Systems, vol. 1, issue 1, (2011).
3. Ye, N., Gauch, S., Wang, Q., Luong, H.:An Adaptive Ontology based Hierarchical Browsing System for CiteSeerX. In: Second International Conference on Knowledge and Systems Engineering (KSE), pp. 203–208, IEEE, (2010).
4. *Wilson, H., Boots, B., Millward, A. A.: A Comparison of Hierarchical and Partitional* Clustering Techniques for Multispectral Image Classification. vol.3, pp. 1624-1626, (2002).
5. Xu, Y.: Hybrid clustering with application to web mining. In: Proceedings of the International Conference on Active Media Technology (AMT 2005), pp. 574–578,IEEE, (2005).
6. Aliguliyev, R. M.: Clustering of Document Collection- A Weighted Approach. In: Expert Systems with Applications,vol. 36, issue 4, pp. 7904–7916,Elsevier, (2009).
7. Boley, D.: Principal Direction Divisive Partitioning. In: Data Mining and Knowledge Discovery, vol. 2, issue. 4, pp. 325 – 344, ACM, (1998).
8. Feng, L., Qiu, M.H., Wang, Y.X., Xiang, Q.L., Yang, Y.F., Liu, K. A.: Fast Divisive Clustering Algorithm Using an Improved Discrete Particle Swarm Optimizer. In:Pattern Recognition Letters, vol. 31, issue. 11, pp. 1216-1225,Elsevier, (2010).
9. Rana, S., Jasola,S., Kumar,R.: A Hybrid Sequential Approach for Data Clustering using K-means and Particle Swarm Optimization Algorithm. In: International Journal of Engineering, Science and Technology,vol. 2, No. 6, pp. 167-176, (2010).
10. *Bache, K., Lichman, M.: UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].* Irvine, CA: University of California, School of Information and Computer Science,(2013).
11. Yang,X. S.: Nature-inspired Metaheuristic Algorithms, 2nd ed., Luniver press, United Kingdom, (2011).
12. Horng, M. H., Jiang, T. W.: Multilevel Image Thresholding Selection based on theFirefly Algorithm. In: 7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC), pp. 58 – 63, IEEE, (2010).
13. Senthilnath, J., Omkar, S. N., Mani, V.: Clustering Using Firefly Algorithm: Performance Study. In: Swarm and Evolutionary Computation, vol. 1, issue. 3, pp. 164-171, Elsevier, (2011).
14. Hassanzadeh, T., Meybodi, M. R.:A New Hybrid Approach for Data Clustering Using Firefly Algorithm and K-means. In: 16thIEEECSI International Symposium on Artificial Intelligence and Signal Processing (AISP), pp. 007 – 011, (2012).
15. Abshouri, A. A., Bakhtiary,A.: A New Clustering Method Based on Firefly and KHM. In: *Journal of Communication and Computer, vol. 9, pp. 387-391, (2012).*
16. Xu, G., Zhang,Y., Li, L.: Web mining and social networking, Techniques and application, New York, Springer, (2011).
17. Manning, C. D., Raghavan,P., Schütze,H.: Introduction to Information Retrieval, 1 ed., Cambridge University Press, (2008).
18. Lewis,D.: The reuters-21578 text categorizationtest collection, 1999.[Online].Available:http://kdd.ics.uci.edu/database/reuters21578/reuters21578.html.
19. Murugesan, K, Zhang,J.: Hybrid Bisect K-means Clustering Algorthim. In: IEEE International Conference on Business Computing and Global Informatization (BCGIN), pp. 216 – 219, IEEE, (2011).
20. Meghabghab, G., Kandel, A.: Search Engines ,Link Analysis ,and User's Web Behaviour, Berlin Heidelberg: Springer-Verlag, (2008).